Perturbations on the uniform distribution of p-values can lead to misleading inferences from null-hypothesis testing

László Zsolt Garamszegi[*] and Pierre de Villemereuil[**]

[*]Department of Evolutionary Ecology, Estación Biológica de Doñana – CSIC , c/ Americo Vespucio, s/n, 41092 Seville (Spain)
[**]School of Biological Sciences, The University of Auckland, Auckland (New Zealand)
Phone: (+64) 9 373 7999, E-mail: bonamy@horus.ens.fr

Abstract

Null-hypothesis testing (NHT) based on statistical significance is the most conventional statistical framework, on which neuroscientists rely for the analysis of their data. However, this approach can provide misleading results if *p*-values are wrongly interpreted, as often done in practice. Misconceptions can arise, in particular, when i) wrong null-hypothesis is chosen for reference; ii) the assumptions of the statistical model are not met; iii) *p*-values are interpreted as the probability of the null- or alternative hypotheses or as the measure of the importance of findings; iv) statistical thresholds guide scientific conclusions and decision making; v) one applies multiple testing or p-hacking. In this commentary, we address these issues by bringing into the focus the uniform distribution of *p*-values with the hope of enhancing the appreciation and proper use of the NHT approach among neuroscientists. We propose guidelines for the correct interpretations of *p*-values that brain and behavioral scientists may adopt to improve both the transparency of statistical reports and the value of scientific conclusions drawn from them.

Keywords: *p*-values, null-hypothesis testing, *p*-hacking, statistics
Published in: Trends in Neuroscience and Education
DOI: 10.1016/j.tine.2017.10.001
Accepted version after peer-review.

Introduction

This paper is a commentary on the recently published formal statements of the American Statistical Association (ASA) about the principles of the proper use and interpretation of *p*-values (Wasserstein and Lazar, 2016). Neuroscientists heavily rely on null-hypothesis testing (NHT), as most empirical studies use significance thresholds to make inferences from the data (Nakagawa and Hauber, 2011). Given that this field of research can be characterized by specific circumstances for the statistical analyses (e.g. hypothesis-driven experimental approach, limited sample size), we believe that the ASA statement deserves amendments from the perspective of this community. Here, we will go through the definition of *p*-values and the statements and comment them to help the objective integration of the principles into neuroscientists' analytical practice. This discussion will focus on the uniform distribution of *p*-values, a chief attribute of the NHT approach that provides hindsight on misconceptions and misuses in hypothesis testing. Finally, we formalise recommendations to improve the practice of the use and report of NHT-based statistics for scientific transparency.

What is the *p*-value?

The ASA statement defines *p*-values as follow:

> [A] *p*-value is the probability under a specified statistical model that a statistical summary of the data [...] would be equal to or more extreme than its observed value.

The textbook example that illustrates the meaning of this definition relies on the comparison of the means of two samples with Student's test. The details of the test are unimportant here, but typically one computes the difference between the sample's mean, then divides it by an estimate of the standard-deviation and obtain a test statistic typically noted t_{obs}. If we assume a particular model, referred to as *null-hypothesis*, we can prove that t_{obs} follows a particular distribution (here, a Student distribution). Hence, we know (still, if this null-hypothesis is true) what is the probability of observing t_{obs} or any value more extreme than that. This is how a *p*-value is computed (see Fig. 1):

$$p = P(T \geq t_{\text{obs}}) \tag{1}$$

assuming T is a random variable following Student's distribution. Note that we are assuming here a one-tailed test for the sake of simplicity, but everything described in this paper applies to both one- and two-tailed tests.

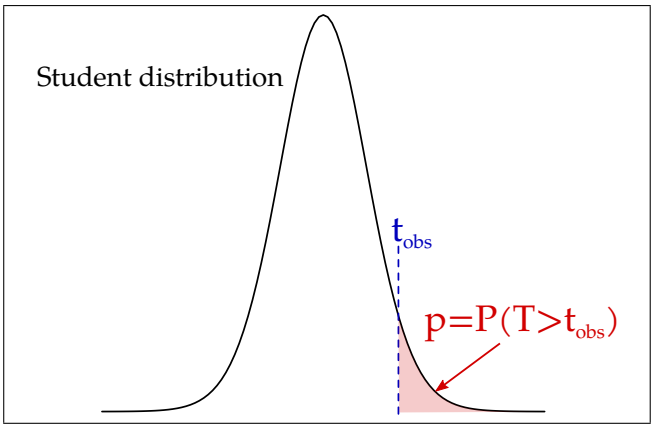

Figure 1: Relationship between an observed statistics (t_{obs}) and its corresponding p-value, here illustrated assuming a t-test and a Student distribution.

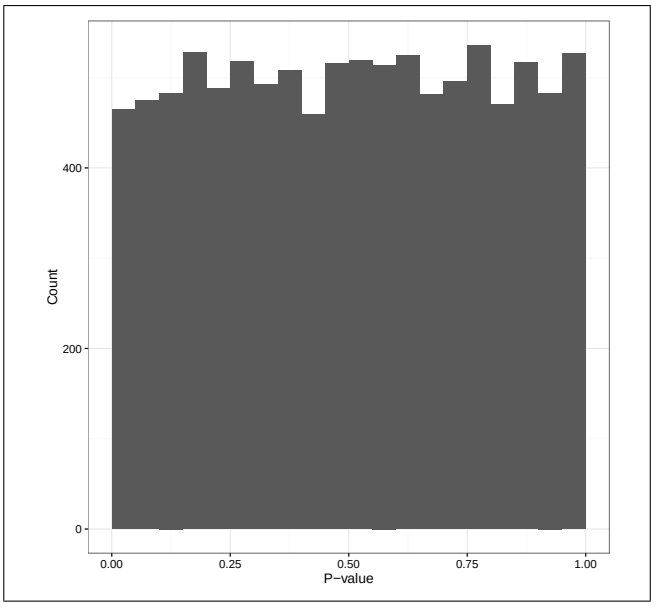

Figure 2: Distribution of the p-values yielded by 10,000 Student test where the two samples are drawn from the same distribution, i.e. the null-hypothesis is true.

A feature of p-values, which we will refer throughout as the "fundamental feature", is their uniform distribution. This feature originates directly from their definition, but is not very intuitive to deduce. However, this feature is completely central to understand why p-values are used for NHT and many of the misconceptions mentioned by the ASA.

Ceci est un test

To see this feature, let's rewrite the definition of p:

$$p = P(T \geq t_{\text{obs}}) = 1 - P(T < t_{\text{obs}}) \tag{2}$$

Note that $P(T < t_{\text{obs}})$ is the cumulative probability corresponding to the quantile t_{obs}, meaning a proportion $P(T < t_{\text{obs}})$ of the Student distribution is below t_{obs}. For continuous distributions, there is an equivalence between a quantile and its cumulative probability, meaning that choosing a probability from a uniform distribution to compute the corresponding quantile is equivalent to drawing directly from the distribution. The direct consequence is this: under the null-hypothesis, the p-values are distributed according to a uniform distribution between 0 and 1. Fig. 2 shows the uniform distribution of p-values of 10,000 repetitions of a Student test for two samples drawn from a standard Normal distribution (hence the null-hypothesis is true). Note that, here, the default two-tailed test was used.

The fact that, under the null-hypothesis, the p-values are uniformly distributed yields to the following property for any given α:

$$P(p \leq \alpha) = \alpha \tag{3}$$

In other words, for any given threshold α, the probability, if the null-hypothesis is true, that the test yields a p-value lower than this threshold is α. This explains the relationship between p-values and the false positive rate (FPR) of a test, i.e. the probability of rejecting the null-hypothesis when it is true. Indeed choosing a significance threshold of 0.05 for the p-values means that, if the null-hypothesis is true, we will be wrong 5% of the time. Appreciating the FPR (a.k.a. Type-I error) is extremely important for the scientific interpretation of the outcome of the NHT because it implies that statistically significant results are always loaded with a given amount of uncertainty concerning the justification of the rejection of the null-hypothesis. For the relationship between p-values and FPR to hold, it is thus necessary that the p-values are uniformly distributed between 0 and 1. *Anything* that will result in a disturbance of this distribution will thus result in an illusory FPR control, which we will highlight in the following sections focusing on particular ASA statements.

Statements

1. *P*-values can indicate how incompatible the data are with a specified statistical model. This statement means that p-values describe the probability of observing a given data or any more "extreme" data if a null-hypothesis is true (this is simply another way to formulate the definition given above). Therefore, p-values describe a *conditional* relationship between the data at hand and the underlying null-hypothesis. The consequence of this conditionality is that p-values do not say anything about the compatibility of the observed data with alternative hypotheses/models (which might as well be wrong). Therefore, a low p-value can be a motivation for disfavouring the considered null-hypothesis as an explanation for our data, but it cannot be used as a support for any other hypotheses. In other words, if a wrong null-hypothesis is chosen as a reference, it will systematically produce very low p-values but without having any implication for other alternatives. Fig. 3 depicts some scenarios when the compatibility of the same data is assessed against different null

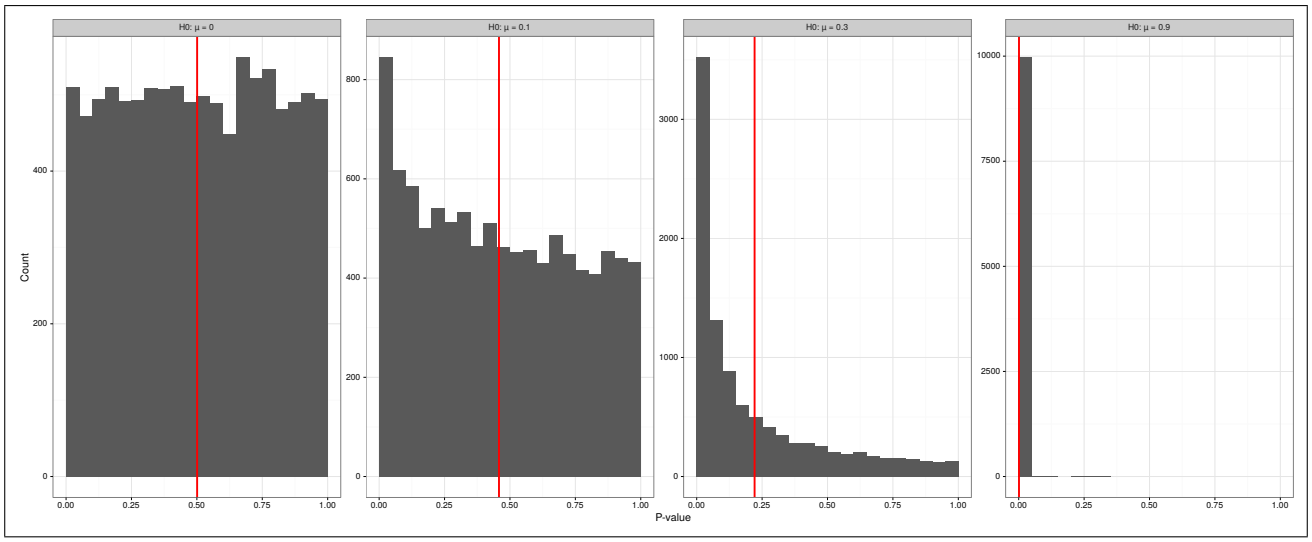

Figure 3: Distribution of the p-values of a t-test of the same data when tested against different hypothetical means.

models. For this figure, we simulated data under a model that assumes a normal distribution with a zero mean and unit variance. Then we challenged the generated data with series of one-sample t-tests that differ in the hypothetical mean values they rely on. What is evident from this exercise is that the uniform distribution of p-values does not hold when the model used for data simulation is different than the null-model considered in the statistical analysis of the generated data. Since we simulated the data for illustrative purposes, in this example we know that the chosen null-hypothesis is true or not. However, this is not the case in most of the empirical situations, when we have no clue about the probability that the studied hypothesis is true. Typically, the null-hypothesis is a statement about the absence of an effect, i.e. there is no difference between groups, or no relationship between two variables. Yet, biologically such null models sometimes embody nonsensical situations, when small p-values only indicate that a silly null-hypothesis was considered for the evaluation of the data (Cherry, 1998; Johnson, 1999; Anderson et al, 2000; Guthery et al, 2001). For the interpretation of p-values, hence, it is always essential to understand that it is always specific to the chosen null, and that the NHT output only tells us how our data fit with this reference but nothing more.

One important component of the first ASA statement is that for p-values to be interpretable the underlying assumptions of the model should be held. This is for a good reason: the uniform distribution of p-values is warranted only when the null-hypothesis is true, which includes those assumptions. If those assumptions are not met, then the null-hypothesis of no effect might seem highly incompatible with the data even in the absence of the effect, simply because it is not true. To illustrate this, let us modify the t-test example. Now, one sample comes from a Normal distribution, but the other one is sampled from a χ^2 distribution with 1 degree of freedom. Both of those distributions are set to share the same mean of 1 (no "biological" effect). The resulting distribution is shown in Fig. 4: we can see that the distribution of p-values is heavily altered by the violation of the Normal assumption made by the t-test, despite there being no change in average value between the two samples. As a consequence, a 5% threshold yields a FPR twice as high as expected (9.5% of positive tests). As another example, we can envisage the situation, when the residuals of a regression model are not independent (e.g. because of spatial/temporal autocorrelation or phylogenetic relatedness). If the data are analysed with a conventional linear model that assumes normal, homoscedastic and independent residuals, the probability of the null-hypothesis of no relationship between variables becomes very small. However, if the same data are submitted to a

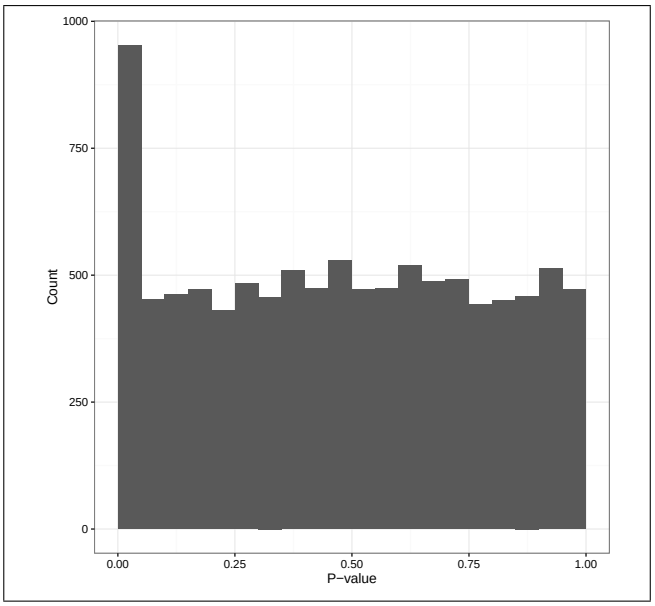

Figure 4: Distribution of the p-values of a t-test when one of the sample is not normally distributed.

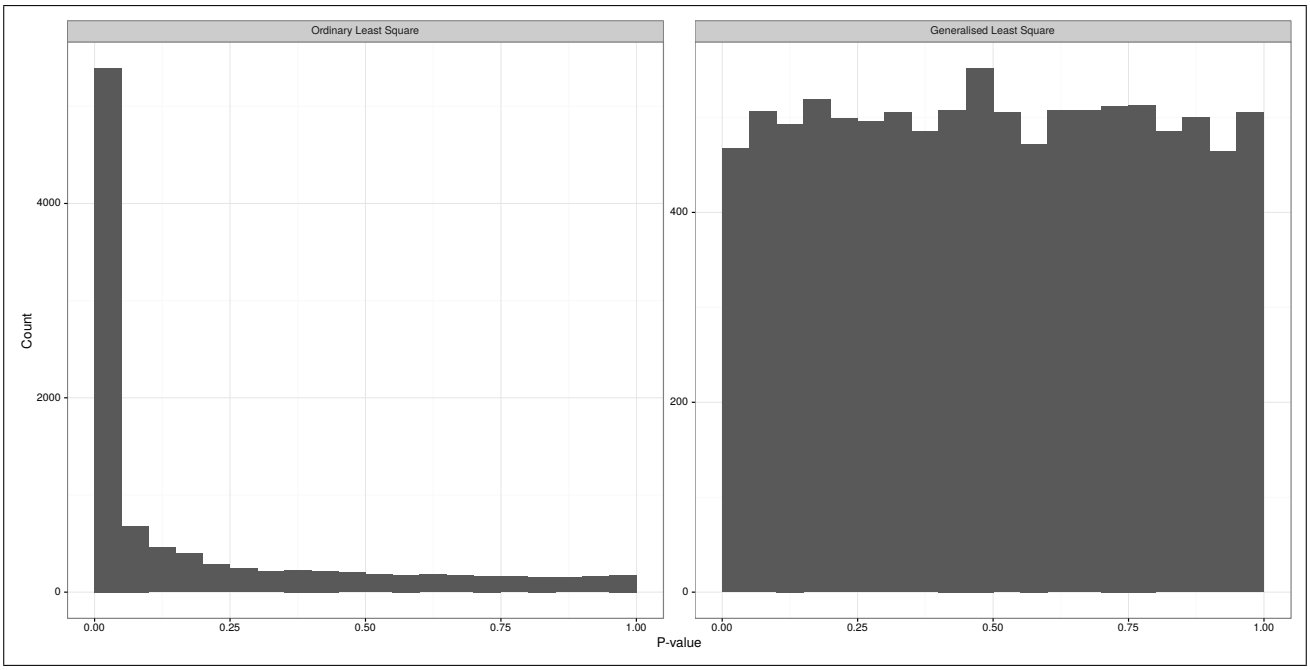

Figure 5: Distribution of the p-values from a regression that assumes non-independent residuals (Ordinary Least Squares) or accounting the correlation structure (Generalised Least Squares) while these are autocorrelated.

model that incorporates the appropriate dependence structure, the uniform distribution of p-values can be regained (Fig. 5). Consequently, the assumptions are the key components of the null-hypothesis that are need to be checked before interpreting any significance test.

2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. There are at least two logical flaws with interpreting p-values as being the probability of the null-hypothesis. First, the definition of a p-value already assumes that the null-hypothesis is true, so it would be nonsensical that p then measures the probability of the null-hypothesis. Second, the fundamental feature states that if the null-hypothesis is true, any value of p is equally likely, meaning that p itself would be a silly measure of this probability. Fig. 3 nicely illustrates these points. From the underlying simulation, we know that the probability of the null-hypothesis is 1 for the figure in the left, because the data were specifically generated under this model. In spite of this, p-values can take any values between 0 and 1 with equal probability. In the other figures, the probability of the null-hypothesis is 0, because we know that we used different models for data generation. In these cases, p-values tend to be lower than 1 and their distribution is skewed towards 0, but none of the observed p-values are exactly 0 (in the whole simulation, the lowest p-value that we observed was 8×10^{-12}). These figures also show the falsehood of equating p-values with the probability that the data were produced by random chance alone, since for each figure the data were generated by random chance alone. The probability of random chance is, therefore, 1, but p-values we obtained varied within the range between 0 and 1 (in the whole simulation, the highest p-value that we observed was 0.9999639).

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold. Given their definition and their fundamental feature, p-values are a useful tool for significance testing, because, by using a pre-established significance threshold, it allows for a transparent control of FPR, i.e. the probability of accidentally finding effects when there are none. However, it is important to note that p-values are informative regarding the error we might commit while rejecting the null-hypothesis, and *only that*. In fact, they are not directly informative about this error, only through they long-run distribution (to which we most often don't have access in practice). From the days of Fisher (1925), the most commonly used threshold for such an error rate is $\alpha = 0.05$. Since then, "we teach it because it's what we do; we do it because it's what we teach." (Wasserstein and Lazar, 2016). However, this is a completely arbitrary threshold that is chosen for some convenience and is independent from any biological motivation. Therefore, it is also completely sound scientifically to establish a predefined (and clearly stated) threshold, say, at $\alpha = 0.0743$ or $\alpha = 0.0214$, and appreciating such risk for mistakenly rejecting the null if it's true. In terms of evidence for rejecting the null-hypothesis, the difference between these scenarios is not important. Applying threshold to p-value is always good for FPR control, but it does create an artificial asymmetry for scientific arguments about the existence or absence of biological effects. Whereas deciding if an effect is "significant" with a yes-or-no response might be needed for policy or business decision makers on the one hand (i.e.

about commercialising a new treatment), one might argue that most researchers, on the other hand, have the luxury (even the duty) of acknowledging uncertainty about their hypotheses.

Possibly the most important problem with using a threshold based on p-values for justifying scientific claims is associated with the issues of ignoring effect size, as discussed under statements 5 and 6 (see below). Particularly, since p-values measure neither the size of the effects nor the importance of findings, and do not embody a reliable measure of evidence, they cannot serve as a strong motivation for concluding that a hypothesis is "true" or "false". Imagine for example that a very well conducted analysis on the relationship between broccoli consumption and Alzheimer disease find (after carefully accounting for any confounding effect) an increased risk with a p-value of 6.3×10^{-8}. Would that be enough to justify a ban on broccolis? Of course, replications of the study would be needed, as well as possible confirmation on animal models and study of possible mechanisms before having a ban in place. But even investing money in these studies is a decision we might not want to make, for a very good reason: the p-value does not tell us *how much* the risk is increased. Would time, money, death of animals and change in our diet be justified for a very significant increase of 0.01% of the risk of Alzheimer disease? Most probably not.

For further discussion about the shortcomings of the use thresholds to mediate "binary" thinking about effects that operate in a continuous scale in nature see Stephens et al (2007b), Hurlbert and Lombardi (2009) or Murtaugh (2014).

4. Proper inference requires full reporting and transparency The definition and uniform distribution of p-values imply that if we perform a large number of tests on data that are completely compatible with a null-hypothesis, in some instances we can obtain very small p-values (Fig. 2). If we work with the conventional FPR at $\alpha = 0.05$, one out of twenty tests will be significant and suggest that the null-hypothesis can be rejected even if it is true. Given that such an error rate is inherent to NHT, one should be extremely careful when conducting multiple tests on the same data. In particular, performing multiple analyses and reporting only those that surpass a particular significance threshold leads to uninterpretable p-values and wrong scientific practice. For example, if we put a dead salmon into an fMRI scan and take a large number of images, some of these will reveal a significant difference between signal and noise levels by chance (i.e. false positives) leading to nonsensical conclusions about brain activity when statistical thresholds are left uncorrected (Bennett et al, 2011). We will investigate two further scenarios here, p-hacking and hidden multiple testing that, based on our experience, we feel important to bring into the attention of practising neuroscientists.

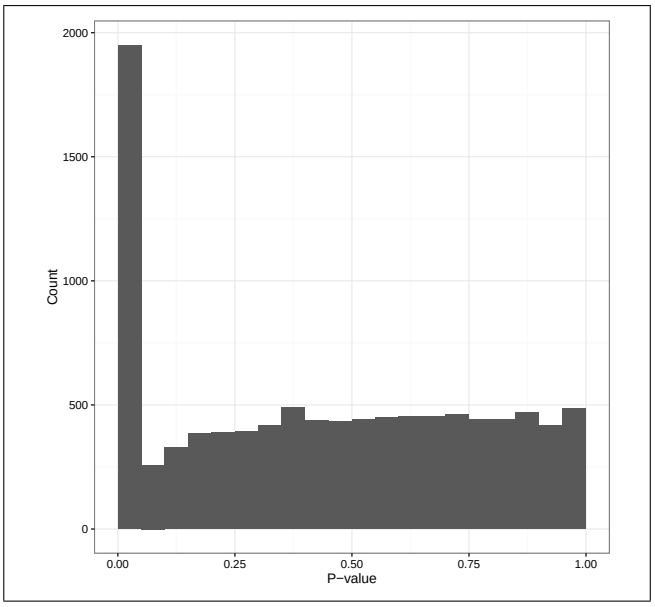

Figure 6: Distribution of the p-values in absence of any effect when p-hacking is used.

First, let us imagine a caricature of a common situation, in which a researcher is assessing some forms of brain activity under two experimental conditions (say a control and a stimulated/inhibited treatment). Because such studies need to be carefully executed, it takes a long time, and our researcher can only assess 5 individuals in both conditions a week. Entering the data and analyse them is much quicker, so the eager researcher computes the test at the end of each week and obtain a p-value. If the p-value gets below 5%, the researcher decides that s/he can stop the experiment because s/he has enough data to say the effect is real. If, after 10 weeks of behaviour assessment (sample size of 50), the test is still negative, the researcher gives up and deems the result of the experiment as negative. This might sound like as sound approach, but it isn't. It is actually a form of p-hacking (Ioannidis, 2005; Forstmeier et al, 2016) and has a strong impact on the distribution of p-values. Fig. 6 shows again a simulation of this exact behaviour and the resulting distribution of p-values in the absence of any effect: the impact is very important and it results into a FPR of 19.5% at a 5% threshold (four times as high as expected). The above example assumes a naive researcher, but even worse, p-hacking can also occur intentionally when one performs several tests at the same time and then aims at publishing the outcome from those that were associated with significant p-values. Such a practice can lead to "scientific" results that show that psychologist students can see future (Galak et al, 2012). Statistical tools are now available that can assess the degree of p-hacking occurring in the published literature (Simonsohn et al, 2014), which reveal that p-hacking is widespread in both biological and medical/health sciences (Head et al, 2015).

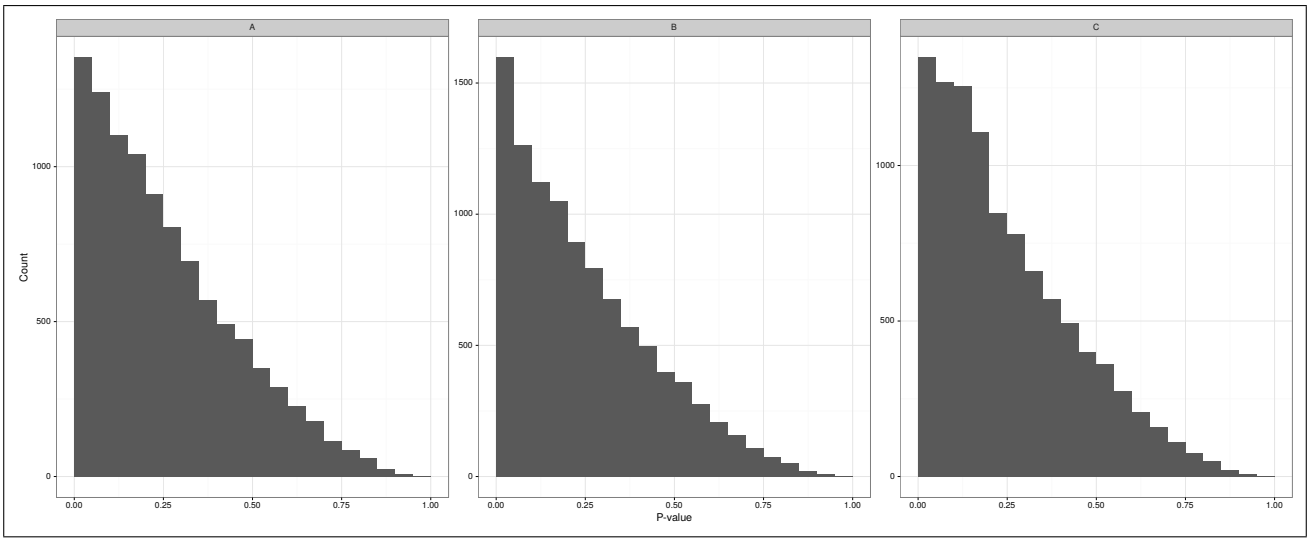

Figure 7: Distribution of the p-values in absence of any effect as consequence of hidden multiple testing. Three sets of independent, normally distributed random data were simulated and entered into a multiple regression model as predictors. Their effects on a fourth randomly generated response variable were assessed in a 10,000 series of repetitions relying on a sample size of 30 for each variable. Panel A shows the distribution of the smallest p-values from these models. Panel B is obtained when a stepwise removal of terms applied until only terms with significant p-values or the one with the lowest p-value remained in the final model. Panel C shows results when the stepwise procedures were based on AIC

Second, let us now envisage an illusory situation, in which a researcher has an access to a system that can measure neuronal activity in several brain nuclei, and the researcher wishes to study their relationships with a certain behavioural trait. We assume that the researcher knows that testing the correlations between the behaviour and the activity of brain nuclei one by one and selectively publish only the significant findings that comes out from this fishing exercise is a bad practice. S/he would like to perform a single analysis instead, and decides that s/he enters the physiological variables into the same multiple regression as predictors, and examines their slopes on the behavioural variable from the fitted model. Then the estimated parameters (slopes and intercept) and the associated p-values are printed into a result table. Until this point, this is a sound approach, but if significant p-values from the table are interpreted as supports at 5% FPR for the rejection of the null hypotheses for the particular variables being unrelated to the behaviour, the researcher enters into the trap of hidden multiple testing (Forstmeier and Schielzeth, 2011). Fig. 7 demonstrates how the distribution of p-values is affected when three randomly generated variables entered into a multiple regression model that tests for their relationship with a random response variable. Panel A corresponds to the situation when the variable with the lowest p-value from the multiple regression is considered only (the approach that is followed by the above researcher who selectively wishes to focus on the significant effects). In terms of FPR, it means that out of 100 multiple regressions based on completely independent variables, there will be on average 13.5 that reveal at least one significant slope just by chance. Panel B is for the scenario when a stepwise procedure based on the removal of the least significant effects is applied until only significant terms remain in the model. The stepwise approach typically assumes that terms not included in the final model have zero effects. Hence, if all variables are removed until the final step they are considered to fulfil the underlying null-model, thus we could assign a p=1 to each of them. For illustrative purposes, in our exercise we kept the variable with the lowest p-value in the final model, even if it did not pass the significance threshold (to avoid the preponderance of values at p=1 of the histogram). Finally, panel C shows the results that are based on a stepwise procedure relying on Akaike's information criterion (AIC) for the removal of terms instead of p-values (which is actually a bad practice but still used frequently, see Whittingham et al, 2006). Each case illustrates that hidden-multiple testing generates perturbations on the uniform distribution of p-values, and it is misleading to selectively focus on the significant p-values in a multiple regression. This is because, the significance of terms in a multiple regression corresponds to different null-hypotheses (the particular term is zero), thus the number of significance tests applied is equal with the number of terms in the model, even if we are considering a single model. Note that this argument applies for the significance of particular variables in the model, while the significance of the full model including all effects is not a concern here. This is because, the p-value for the full model refers to the compatibility with a single null-hypothesis (the only intercept model). Note that a problem akin to multiple comparisons can appear even when a unique test is performed with no apparent "p-hacking", due to the researcher designing statistical tests based on how the data at hand are generally behaving (so-called "researcher degrees of freedom" Gelman and Loken, 2013).

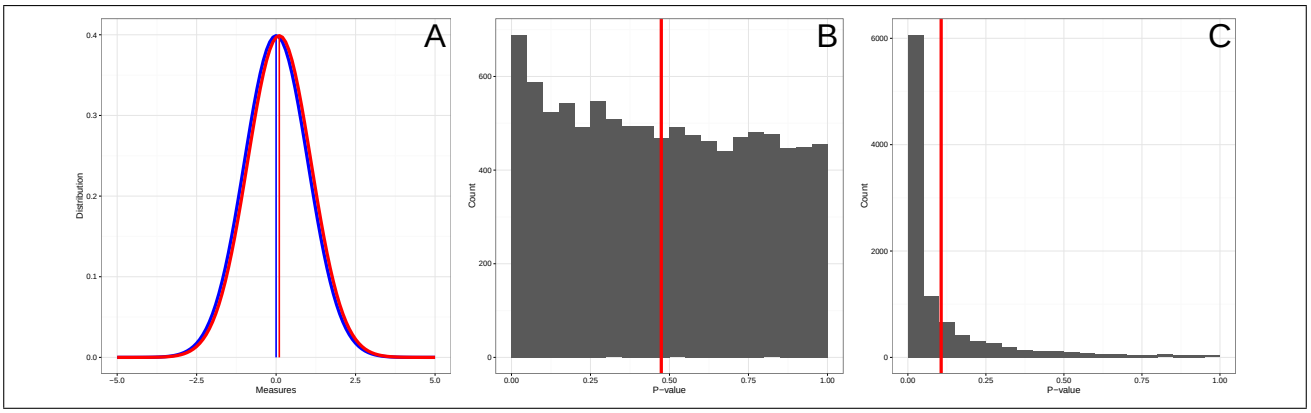

Figure 8: The p-values distribution for testing a small effect. A model description of the test in on A: two populations (blue and red) are testing and differ in average by a very small amount (mean depicted as vertical lines). On the right, distribution of the p-values when 10,000 test are performed with a sample size of 30 (B) or 1,000 (C). Mean p-value is depicted as a red vertical line.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. When the null-hypothesis is false, it might get rejected because of a low p-value. Because the lowest the p-value is, the strongest is the case against the null-hypothesis, one might conclude that a lower p-value is the sign of a stronger effect. But this is not the case, as the relationship between the p-value and the effect size is rendered complicated by many factors such as the level of noise in the data (e.g. due to biological noise or measurement error) and the sample size (e.g. a test might be significant over a tiny effect, provided the sample size is big enough; on the other hand biologically important effects might be associated with non-significant p-values if sample size is small). To illustrate this, let's consider a Student test performed with a very small effect size. Fig. 8A shows the distribution of two populations differing in 0.1 in their average value. We compare 10,000 simulated tests with a sample size of 30 or 1,000. With a small sample size, we observe only a slight deviation from the uniform distribution (Fig. 8B). This results in an average significant (at the 5% threshold) p-value of 0.023 (the average of significant p-values expected under the null-hypothesis is $0.05/2 = 0.025$). With a large sample size however, the uniform distribution is very distorted, with a very high peak of small p-values (Fig 8C). Here, the average significant p-value is 0.012. Since both simulations have the exact same biological settings (e.g. same effect size) and differ only by their sample size, it is obvious from this toy example that the p-values are a bad estimate of the biological effect size.

Furthermore, statistical significance provides no measure for either the sign or the magnitude of the effect being correctly estimated. In a study that is loaded with measurement errors and constrained by small sample size, there is an increased chance to estimate effects that have a sign in the wrong direction (Type S error) or with an extreme overestimation of the effect size (type M error, Gelman and Loken, 2013; Loken and Gelman, 2017). Therefore, because of filtering of the effect sizes by statistical significance, the significant estimates will be more likely to be of severely biased magnitude. For example, in the above simulation study, when the sample size was low ($N = 30$), 13% of the significant effects had a sign in the wrong direction (type S error, i.e. the mean of the red population was smaller than that of the blue population), while the effect size was on average overestimated by a factor 6 (type M error). The picture was different for the large sample size scenario ($N = 1000$, type S error at 0% and type M error at 1.28).

In fact, sample size, effect size and p-values are all describing different aspects of the statistical testing and should always be reported together. Effect sizes are not known a priori (such as we have no information about the reliability of the null-hypothesis), but these are what we actually intend to estimate in the form of correlations, slopes or differences between experimental groups etc. (or some standardised derivates of these, see Nakagawa and Cuthill, 2007). These are continuous measures, and not a binary state variables describing if there is an effect or not (as the binary thinking would enforce us to think about nature). The precision of these estimates is determined by the sample size and measurement errors, and in fact always should always accompany parameter estimates in the form of standard error or confidence interval. Therefore, if an estimated effect size is associated with low precision (wide confidence range) we will have high probability to observe data that confines the null-hypothesis stating that the focal effect size is zero. Therefore, the effect size captures the biological importance that we wish to estimate, while the precision (or confidence) of this estimate is determined by the constraints of the data at hand. Although both of these aspects contributes to the importance of the results, mean estimates of the effect size and their precision reflect completely different things. Being a summary statistics, p-values combines these properties into a single test statistics. This is handy for the particular process of NHT, but it means that looking at p-value alone cannot allow one (especially a reviewer) to separate the influence of effect size from the effect of sample size.

6. A *p*-value does not provide a good measure of evidence regarding a model or hypothesis This statement can be directly deducted given the uniform distribution of p if the null-hypothesis is true. How could something that can take equally any value between 0 and 1 if the null-hypothesis is true bear any evidence in favour of this hypothesis? Having a p-value of 0.80, for example, tells us little about the evidence in favour of the null-hypothesis apart from the fact that it cannot be ruled out. P-values do bear evidence *against* the null-hypothesis, but only when they are low: the lowest the p-value is, the more "surprising" the result of the test was, should the null-hypothesis be true. However, in this case, we still do not have any evidence on how other alternative models explain the observed data.

Recommendations

Keep your distribution uniform As we saw, p-values, and especially their significance threshold, are only meaningful if their distribution, should no effect be real, is uniform. In practice, one does not have access to this distribution, so how can we ensure this? Here we have pointed to some issues that can cause perturbations on the uniform distribution of p-values under the null-hypothesis. Accordingly, violating model assumptions, p-hacking, hidden multiple testing are obvious scenarios that should be avoided. We feel essential to reiterate that any statistical test have various axillary assumptions (normality, homoscedasticity, etc...), and these assumptions need to be carefully checked before applying a statistical test. This is not only because the deviation of these assumptions can have an influence on the interpretability of p-values, but also because such violations can raise biases in parameter estimates or make the model sensitive to small changes in the data. For further discussion about the importance of model checking see Zuur et al (2010), Loy and Hofmann (2013) or Mundry (2014).

The issue about p-hacking revolves around the realisation that the process of statistical testing needs to be *independent* from data collection. Assessing significance along data collection or sub-sampling data toward significance either intentionally or unintentionally are dangerous for scientific evidence because *(i)* they are heavily hampering the uniform distribution of p-value and *(ii)* they are hardly detected by the reviewing process and can generate publication bias.

When a large number of tests are performed simultaneously p-values are not suited to control for FPR, even if this is done within a single multi-predictor statistical model. Therefore, the interpretation of p-values in the case of multiple testing is challenging. Different correction methods exist that can be used to achieve a reasonable control on FPR. The commonly used Bonferroni correction is often conservative because it focuses the FPR control on the overall analysis, as if one expects only a single test to be positive among all performed tests (hence loosing power as the number of performed tests increases). This constraint is relaxed when using the false discovery rate (FDR) that focuses on the proportion of false positives among the significant results and is based on the whole distribution of p-values (Benjamini and Hochberg, 1995; Storey, 2003; Storey et al, 2004). Therefore, the FDR approach embraces the fundamental feature that p-values from tests with non existing effect must be uniformly distributed (François et al, 2016). Although, we cannot provide an exhaustive and balanced review on the existing correction methods, we note that finding the best correction method for the data and question at hand is often challenging. The issue we want to bring into attention here is that when multiple testing is applied, uncorrected p-values are non-interpretable (Bennett et al, 2009).

We must admit that in our discussion, we may not have identified all scenarios that result in the distortion of the uniform distribution of p-values. However, relying on the flexibility of modern statistical computing environments (such as in the program R), it is relatively easy to perform a simulation study to examine this fundamental feature in association with the current statistical situation. Similarly to the philosophy we followed above, one can generate data under the applicable analytical design and data constraints considering that the underlying null-hypothesis is true, then fit the statistical model and after several repetitions examine the distribution of p-values from the simulation outputs. If this distribution shows patterns of deviance from being uniform, one may conclude that the fundamental condition of NHT is not met, thus p-values from the model that is fitted to the real data may be misleading.

Is the considered null-hypothesis appropriate? One of the chief points of the ASA statements was that a given p-value is always conditional to a particular null-hypothesis. Therefore, if an implausible (i.e. silly) null-hypothesis is chosen for reference, interpretations from an arbitrarily low p-value are deceiving. The plausibility of a null-hypothesis is not always straightforward, as it may depend on prior knowledge, the way of reasoning and the biological question. A typical null-hypothesis posits the absence of an effect, i.e. that there is no difference between two groups, or there is no relationship between two variables (correlation or slope equal to zero). In some cases, the consideration of such null-hypothesis might be nonsensical (e.g. when contrasting a group of mice with a group of elephants, or when investigating the allometric relationship between brain size and body size when the correlation between variables is non-zero by definition). In other instances, plausibility may be context-specific, and a reasonable null-hypothesis can turn into a highly implausible one due to the accumulating evidence. Although based on careful scientific thinking and *biological* motivations sen-

sible null hypotheses that allow proper inferences from *p*-values can be formulated in most of the cases, these might be different than what the default models propose based on *statistical* considerations. For example, the significance that is given by most statistical packages for a complex model including many predictor variables yields the incompatibility of the data with the null model containing only the intercept. However, biologically, a more rational null model would be a model that includes not only the intercept, but also the control variables (with which the full model could be contrasted with a likelihood ratio test). Note that the significance of particular parameter estimate (i.e. a slope or intercept) in the same model reflects a comparison with a null model that includes the same list of predictors except the focal parameter (or in other words, a null model in which the parameter is forced to be zero). It is the researcher's responsibility to check the biological meaning of the null-hypothesis, and eventually smartly redefine a null-hypothesis that does not inherently postulate zero effect. We keep on reminding the readers that null hypotheses should be handled jointly with the auxiliary assumptions, thus the considerations about the plausibility of a null-hypothesis also include issues about these assumptions.

Interpret the *p*-value correctly We re-articulate here that *p*-values cannot tell much in favour of the null-hypothesis, about the strength of the relationships or importance of the results, and about the probability of an alternative hypothesis. Therefore, researchers may want to avoid statements along these lines. The sole and only meaning of a significant *p*-value is that the null-hypothesis might be rejected with a certain level of confidence (typically 5%). This does not necessarily mean that "there is an effect", it might well be that the null-hypothesis is rejected because it was a bad model for the data to begin with (e.g. because an underlying assumption was not met), even in absence of any effect. Yet, these conclusions can only hold at the appreciated FPR indicating that a known proportion (i.e. "alpha") of tests will reject a true null-hypothesis when it is true (type I error or false-positive). A non-significant *p*-value, on the other hand, can indicate that we cannot exclude the null-hypothesis as a potential explanation for the observed data, but this conclusion should not intuitively imply that "there is no effect" and alternative hypotheses cannot be true (see paragraph below about "negative" results). This scenario also incorporates erroneous decisions, as at an unknown rate (i.e. "beta") we will not be able to reject a wrong null-hypothesis though it is false (type II error or false-negative). Both false-positives and false negatives are always integral to hypothesis testing. But it is important to appreciate that 5% FPR does not imply that only 5% of positive results will be false. This latter probability (of false positives among the significant results) is actually called false positive report probability (FPRP) and depends on the FPR (type I error), but also on the type II error and the proportion of true and false hypotheses tested. Typically, and especially at low power or when dubious hypotheses are investigated, the FPRP is much higher than FPR (Ioannidis, 2005; Forstmeier et al, 2016), indicating that a large proportion of significant research findings can be expected to be false.

Correctly interpreting *p*-value significance is one thing, but caution is also recommended when interpreting the estimates of significant models. Not only type I (FPR) and type II (1-power) should be considered when discussed the effect sizes of significant models, but also type S and M errors (Gelman and Tuerlinckx, 2000; Gelman and Carlin, 2014), especially when sample size is small and measurement error is high. These errors describe, respectively, the probability that the sign of a significant effect is in the wrong direction, and the degree by which a significant effect is overestimated. If type S and M errors are high, then we have a high chance to reveal, e.g. a positive/strong relationship when the true relationship is negative/weak. These statistics can be obtained through a design analysis, in which a simulation study is performed, assuming a true effect size, possibly taken from existing literature: data are simulated according to a design identical to the one used in the original study (Gelman and Carlin, 2014), which in turns allow for the computation of type S and M errors.

Transparent reporting The ASA's statement (Wasserstein and Lazar, 2016) stresses that a *p*-value alone cannot be used to draw scientific conclusion, even less policy decisions. First, it should be clear to what null-hypotheses (and assumptions) they correspond. Second, as *p*-values depend on at least two properties, the sample size and the magnitude of the biological effects, these should be reported in parallel. Sample sizes, as shown in Fig. 8, have a strong impact on the ability of the statistical test to detect very small effects. Most often, although the size of the total sample of the study is reported, information on the actual sample sizes used in different statistical tests are missing. The density of missing data might greatly vary among statistical tests relying different sets of variables, it may make comparisons of *p* values across models on different subsamples unreliable. As *p*-values cannot be used directly to formulate statements about the magnitude of the effects, parameter estimates and their standardised derivates can be informative in this regard. Effect size conventions (e.g. sensu Cohen, 1988) can be used to make judgements about the whereabouts of the focal effect along the continuum between small and strong effects (Nakagawa and Cuthill, 2007). However, it should not be forgotten that effect sizes are always estimated with a certain precision, thus they should ideally be supported by the associated confidence interval. In conclusion, to achieve the full transparency of a NHT result, we suggest that the underlying sample size (and/or degrees of freedom if applicable), the estimated effect sizes (standardised or unstandardised) together with its confidence interval are being reported along with the *p*-value. Importantly, it is of vital importance to provide clear picture on the number of hypotheses considered and the

number of statistical tests performed in study. Applying threshold to p-value does tend to put an artificial asymmetry around the threshold. In terms of evidence for rejecting the null-hypothesis, the difference between $p = 0.049$ and $p = 0.051$ is not important. As a consequence, p-values can be commented without a significance threshold, provided the analysis has been performed correctly (i.e. the expected distribution is indeed uniform). In such cases, the p-value can be seen as an estimation of the maximal FPR control which would allow significance. We thus advocate for a systematic reporting of the actual p-values, significant or not, rather than a binned reporting (e.g. $p < 0.001, p < 0.01$ and $p < 0.05$). As p-values that do not surpass the chosen threshold are only informative when type II error rate is convincingly low, analyses of statistical power is pivotal when one aim at retaining a null-hypothesis.

Negative results are interesting There is an ongoing debate regarding negative results (Knight, 2003; Fanelli, 2011; Forstmeier et al, 2016, see also negative-results.org) and publication bias in favour of significant result (Easterbrook et al, 1991; Fanelli, 2010; Franco et al, 2014). This is an issue that reaches beyond the topic of our discussion, but from the NHT point of view, we feel it is important to make a reference to the meaning of a non-significant p-value. First, since under the null-hypothesis, the p-values are uniformly distributed, their value bears little information (but should be reported nonetheless). Second, non-significance simply means that the null-hypothesis cannot be rejected. This might have two explanations: *(i)* the null-hypothesis is indeed true or *(ii)* the null-hypothesis is not true, but the effect size and/or sample are too small for it to be rejected with appropriate confidence. A power analysis can even estimate how small an effect would have to be to be missed by the given analysis. We thus advocate that negative results *are* scientifically informative, provided they are interpreted correctly: they inform us on how small (from non-existent to the maximal effect size the analysis at hand could have missed) an effect could be.

Be aware of alternatives The NHT is not the only framework, in which statistical inferences can be drawn from the data. In fact, sound scientific papers can be written without a single p-value. It is not our purpose here to list all alternatives, but it is important to note that there might be multiple ways for analysing the same data (Nakagawa and Hauber, 2011; Stephens et al, 2007a; Garamszegi et al, 2009). Each of these, such as the one that is based on p-values, can have both strengths and weaknesses. It is pivotal that the researcher understands these characteristics for the correct interpretation of any statistical output. When more than one statistical approach is available for addressing the same biological question, it is often straightforward to apply these in parallel that can be used to evaluate the robustness of the results. If different results are obtained via different methods, the correct interpretation of these differences can be useful to identify the peculiarities of the data or even can be suggestive about biological patterns.

Conclusion

Properly analysing data is as much an important part of the scientific process as the experimental design is. Not realising this state of truth can lead to the general conclusion that "if the experiment was well performed and the p-value is significant, then it must mean something". However, this conclusion is not automatically guaranteed. The only meaning of a $p < 0.05$ is that, at the risk of 5% of being wrong, the chosen null-hypothesis and way of analysing data is not compatible with the observed data, which *might* be due to the presence of an effect, but other information (at the very least, transparent report of the analysis, effect size and sample size) are needed to confirm this. The ASA statements unanimously accentuate that NHT can lead to erroneous scientific process if p-values are misused and misinterpreted. However, these statements do not mean that the NHT-based approach and p-values *per se* are wrong and should be dismissed. We have hope that realising the importance of the fundamental feature of p-values explored here, and how our behaviour as data analyst can destroy this feature, will help our community realise the importance of better practice regarding the use of NHT and the scientific reporting of p-value-based inferences.

Acknowledgement We are grateful to J. Hadfield for useful comments and to N. Cseh for assistance. LZG was supported by funds from The Ministry of Economy and Competitiveness (Spain) (CGL2015-70639-P), and the National Research, Development and Innovation Office (NKFIH, K-115970).

References

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. The Journal of Wildlife Management 64(4):912–923,

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological)

Bennett C, Baird A, Miller M, Wolford G (2011) Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. Journal of Serendipitous and Unexpected Results

Bennett CM, Wolford GL, Miller MB (2009) The principled control of false positives in neuroimaging. Social Cognitive and Affective Neuroscience 4(4):417–422,

Cherry S (1998) Statistical tests in publications of the Wildlife Society. Wildlife Society Bulletin (1973-2006)

Cohen J (1988) Statistical power analysis for the behavioural sciences.

Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR (1991) Publication bias in clinical research. The Lancet 337(8746):867–872,

Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US States Data. PloS one

Fanelli D (2011) Negative results are disappearing from most disciplines and countries. Scientometrics 90(3):891–904,

Fisher RA (1925) Statistical methods for research workers.

Forstmeier W, Schielzeth H (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behavioral Ecology and Sociobiology 65(1):47–55,

Forstmeier W, Wagenmakers EJ, Parker TH (2016) Detecting and avoiding likely false-positive findings – a practical guide. Biological Reviews

Franco A, Malhotra N, Simonovits G (2014) Publication bias in the social sciences: Unlocking the file drawer. Science 345(6203):1502–1505,

François O, Martins H, Caye K, Schoville S (2016) Controlling false discoveries in genome scans for selection. Molecular Ecology 25(2):454–469,

Galak J, LeBoeuf RA, Nelson LD, Simmons JP (2012) Correcting the past: failures to replicate psi. Journal of Personality and Social Psychology 103(6):933–948,

Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jørgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. Behavioral Ecology 20(6):1363–1375,

Gelman A, Carlin J (2014) Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science bibtex: gelman_beyond_2014

Gelman A, Loken E (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Bibtex: gelman_garden_2013

Gelman A, Tuerlinckx F (2000) Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics 15(3):373–390,

Guthery FS, Lusk JJ, Peterson MJ (2001) The fall of the null hypothesis: liabilities and opportunities. The Journal of Wildlife Management 65(3):379–384,

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of p-hacking in science. PLOS Biology 13(3):e1002,106,

Hurlbert SH, Lombardi CM (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46(5):311–349,

Ioannidis JPA (2005) Why most published research findings are false. PLOS Medicine 2(8):e124,

Johnson DH (1999) The insignificance of statistical significance testing. The Journal of Wildlife Management 63(3):763–772,

Knight J (2003) Negative results: Null and void. Nature 422(6932):554–555,

Loken E, Gelman A (2017) Measurement error and the replication crisis. Science 355(6325):584–585, bibtex: loken_measurement_2017

Loy A, Hofmann H (2013) Diagnostic tools for hierarchical linear models. Wiley Interdisciplinary Reviews: Computational Statistics 5(1):48–61,

Mundry R (2014) Statistical issues and assumptions of phylogenetic generalized least squares. In: Garamszegi LZ (ed) Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, Springer Berlin Heidelberg, dOI: 10.1007/978-3-662-43550-2_6

Murtaugh PA (2014) In defense of P-values. Ecology 95(3):611–617,

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82(4):591–605,

Nakagawa S, Hauber ME (2011) Great challenges with few subjects: Statistical strategies for neuroscientists. Neuroscience & Biobehavioral Reviews 35(3):462–473,

Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: a key to the file-drawer. Journal of Experimental Psychology: General

Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2007a) A call for statistical pluralism answered. Journal of Applied Ecology 44(2):461–463,

Stephens PA, Buskirk SW, del Rio CM (2007b) Inference in ecology and evolution. Trends in Ecology & Evolution 22(4):192–197,

Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(1):187–205,

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician 70(2):129–133,

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75(5):1182–1189,

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1(1):3–14,